IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jani Juvani
18.10.2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory Analysis with SQL
  - Exploratory Analysis with Data Visualization
  - Interactive Visual Analytics with Folium lab
  - Interactive Dashboard with Plotly Dash
  - Predictive Analysis by Classification
- Summary of all results
  - Exploratory Analysis results
  - Interactive Analytics results
  - Predictive Analysis results

Introduction

- Project background and context

SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. The goal of this project is to predict if the Falcon 9 first stage will land successfully.

- Problems you want to find answers
  - Which factors determine if the Falcon 9 first stage land successfully?
  - Which classification algorithm predict the Falcon 9 first stage landing success most accurately?

Section 1

Methodology

Methodology

Executive Summary

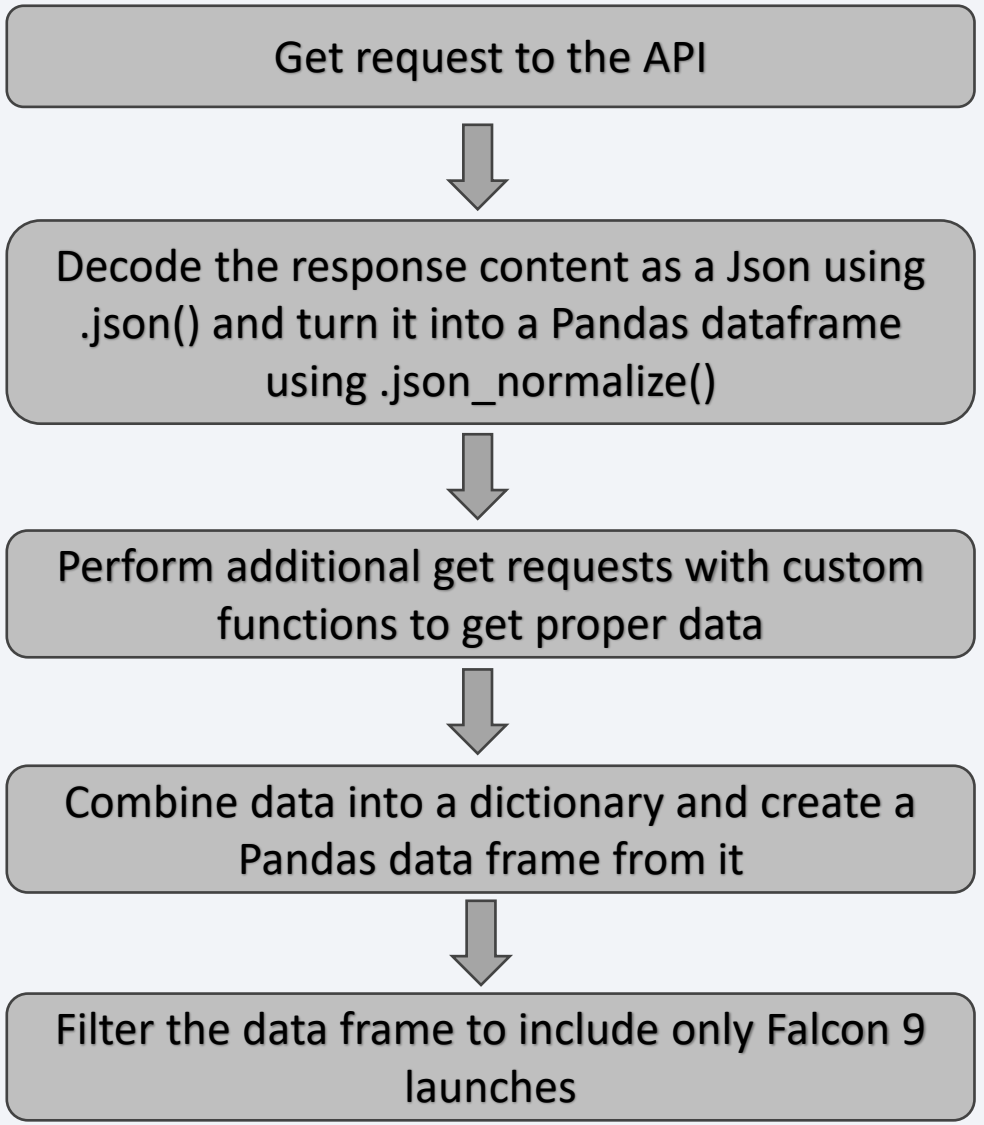
- Data collection methodology:
  - Data was collected using SpaceX API and Web Scraping from Wikipedia
- Perform data wrangling
  - Data was filtered to include only Falcon 9 rocket launches
  - Missing values were replaced by mean values
  - One-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluation of classification models

Data Collection

- SpaceX REST API
  - Data were fetched with get request to the API
  - Raw data were turned into a Pandas data frame
- Wikipedia
  - Data were collected from website: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
  - Web scraping was used to fetch the data
  - Raw data were turned into a Pandas data frame

Data Collection – SpaceX API

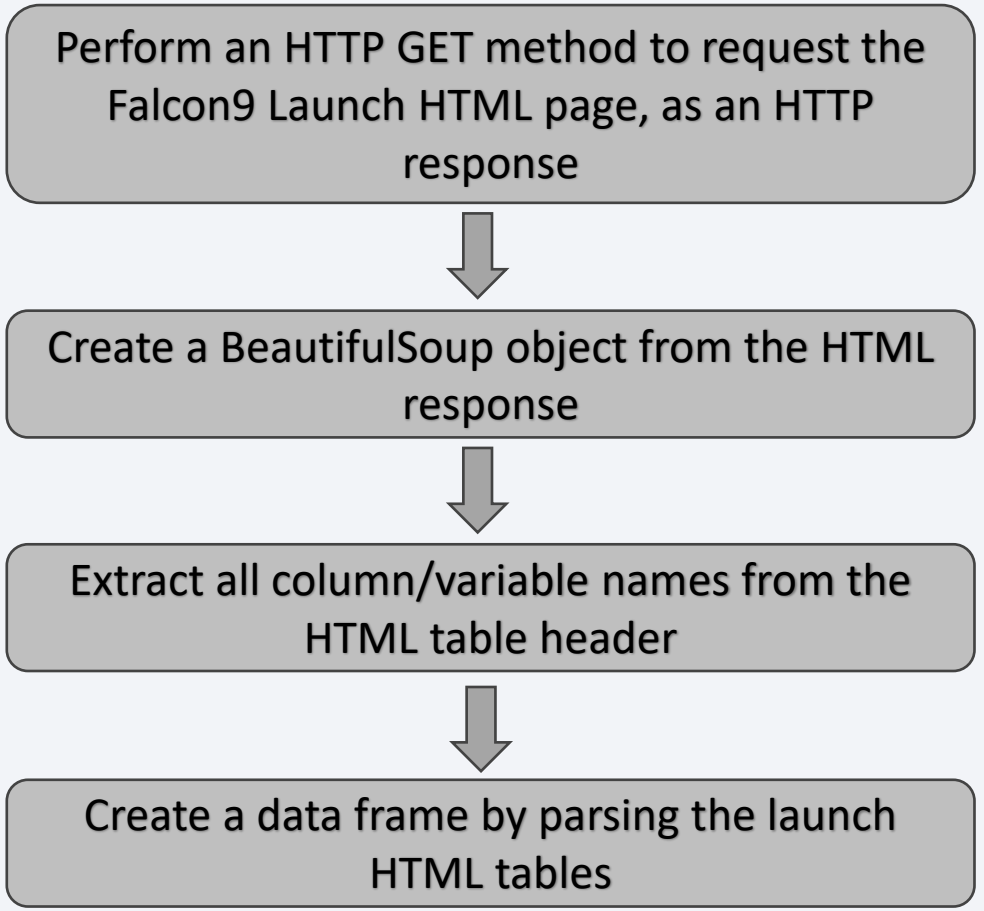
- Data were collected from SpaceX REST API with the protocol showing on the right
- GitHub URL of the completed SpaceX API calls notebook: [Data collection from SpaceX API]



Data Collection - Scraping

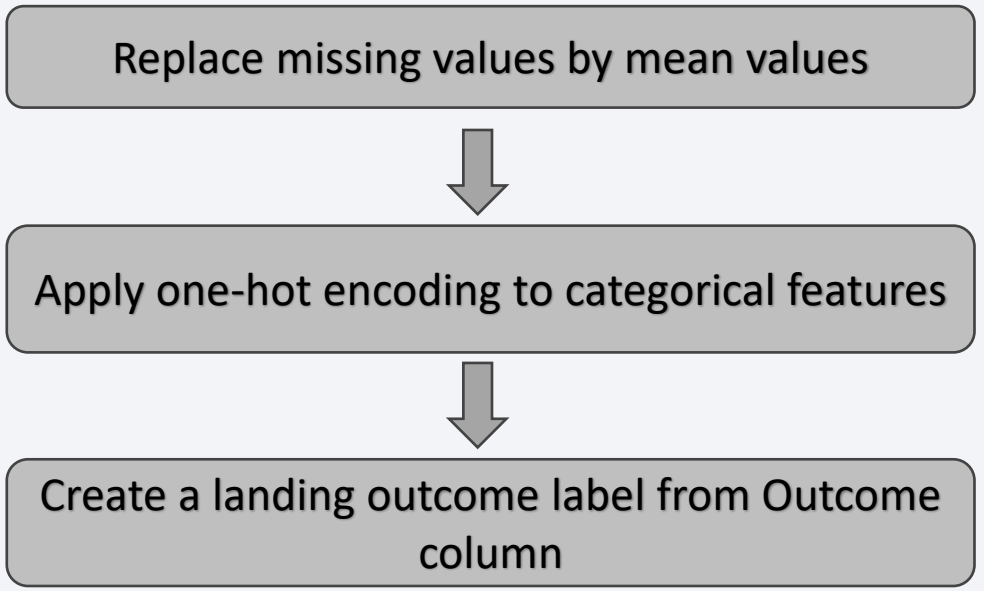
- Data were collected from Wikipedia by web scraping with the protocol showing on the right

- GitHub URL of the completed web scraping notebook: Data collection by web scraping



Data Wrangling

- Data wrangling was performed with the protocol showing on the right
- In the data set, there were several different cases where the booster did or did not land successfully. Those outcomes were converted into training labels with 1 meant the booster landed successfully and 0 meant landing was unsuccessful.
- GitHub URL of the completed data wrangling notebook:
  Data wrangling



EDA with Data Visualization

- Charts plotted:
  - Scatter plot of Flight Number vs. Payload Mass with Success Rate
  - Scatter plot of Flight Number vs. Launch Site with Success Rate
  - Scatter plot of Payload Mass vs. Launch Site with Success Rate
  - Bar plot of Orbit Type with Success Rate
  - Scatter plot of Flight Number vs. Orbit Type with Success Rate
  - Scatter plot of Payload Mass vs. Orbit Type with Success Rate
  - Line plot Launch Success Yearly Trend
- Scatter plots show the relationship between variables, bar plot shows differences between categories and line plot shows the trend over time
- GitHub URL of completed EDA with data visualization notebook:
  EDA with Data Visualization

EDA with SQL

- Performed SQL queries
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster versions which have carried the maximum payload mass. Use a subquery
  - List the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015.
  - Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

Completed EDA with SQL notebook in GitHub

Build an Interactive Map with Folium

- All launch sites were added to map with circles and markers
  - Launch site locations are easy to discover from map and their proximities can be inspected
- Each launch record (successful or not successful) was added to marker clusters
  - Helps to identify which launch sites have good success rates
- Distance between one launch site and coastline was added using marker and polyline
  - To inspect distances between launch site and important landmarks

Completed interactive map with Folium map in GitHub

Build a Dashboard with Plotly Dash

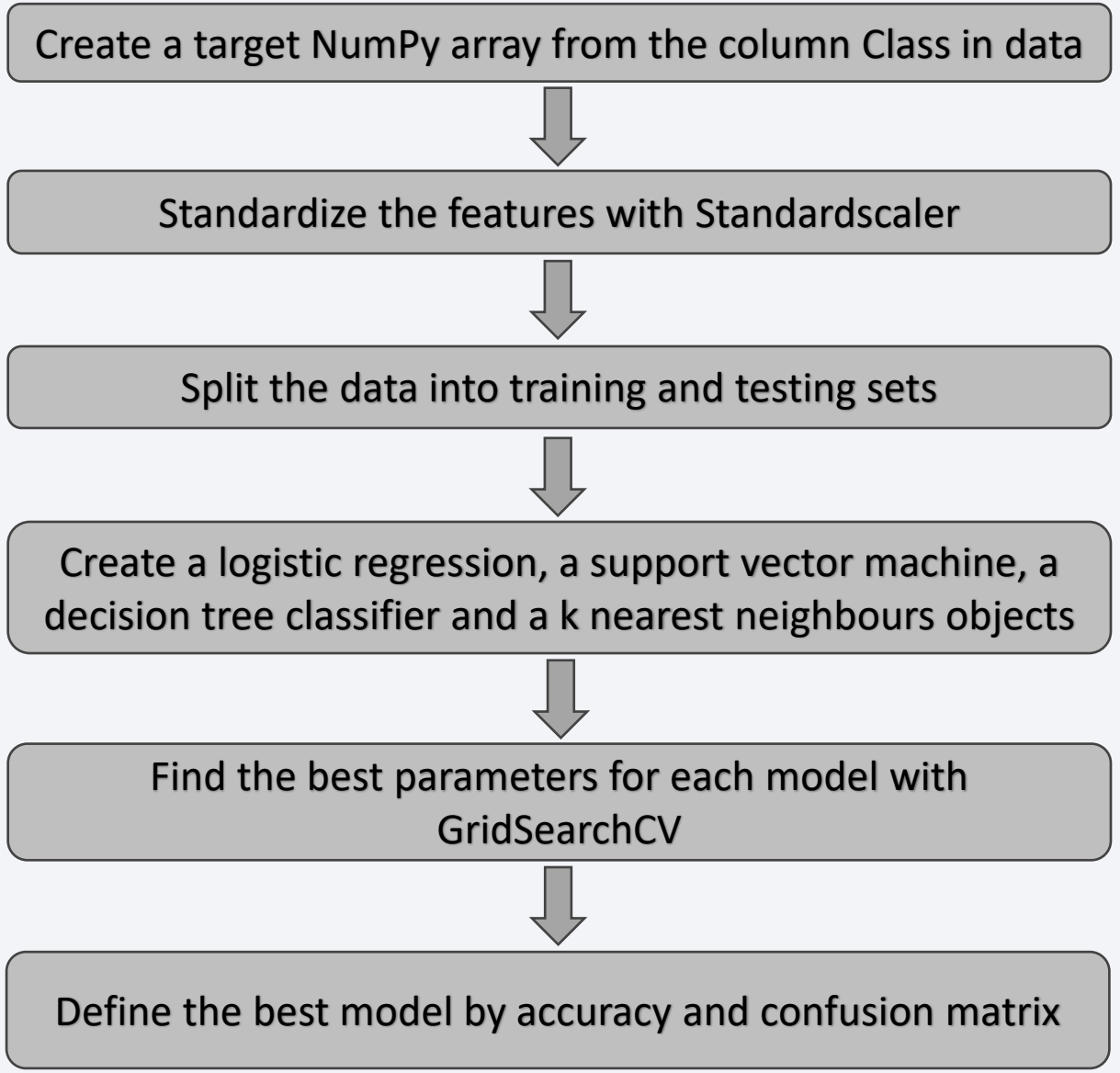
- Launch site dropdown list
  - To enable to select all launch sites or one specific launch site
- A pie chart to show successful launches
  - Total successful launch count for all sites
  - Successful/failed launch ratio for a specific launch site
- A payload range slider
  - To inspect success rate with different payloads
- A scatter chart to show the correlation between payload and launch success
  - To inspect how payload affects the success rate
  - Booster version was also added to the chart

Completed Plotly Dash lab in GitHub

Predictive Analysis (Classification)

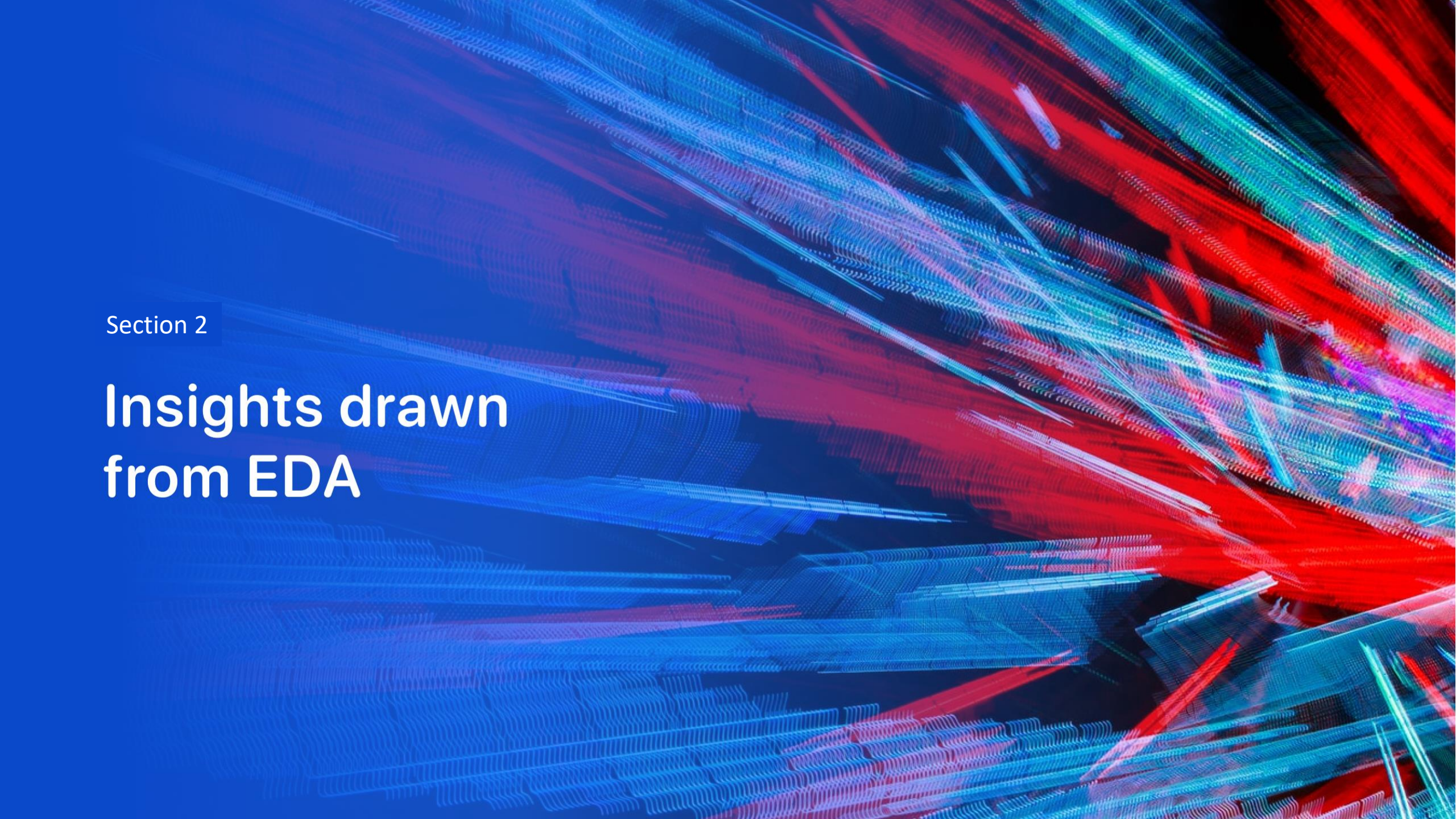
- The best performing classification model was found with the protocol showing on the right

Completed predictive analysis lab in GitHub



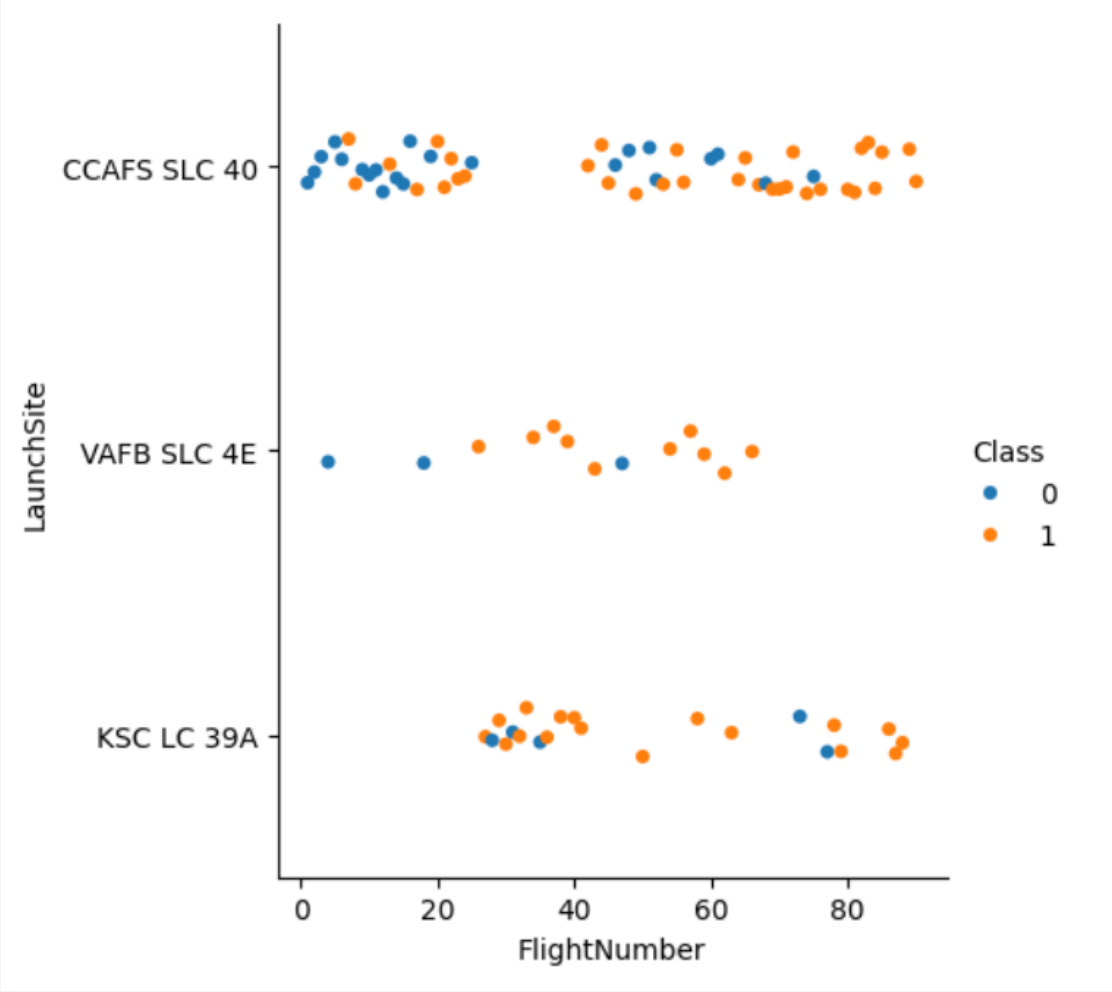
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

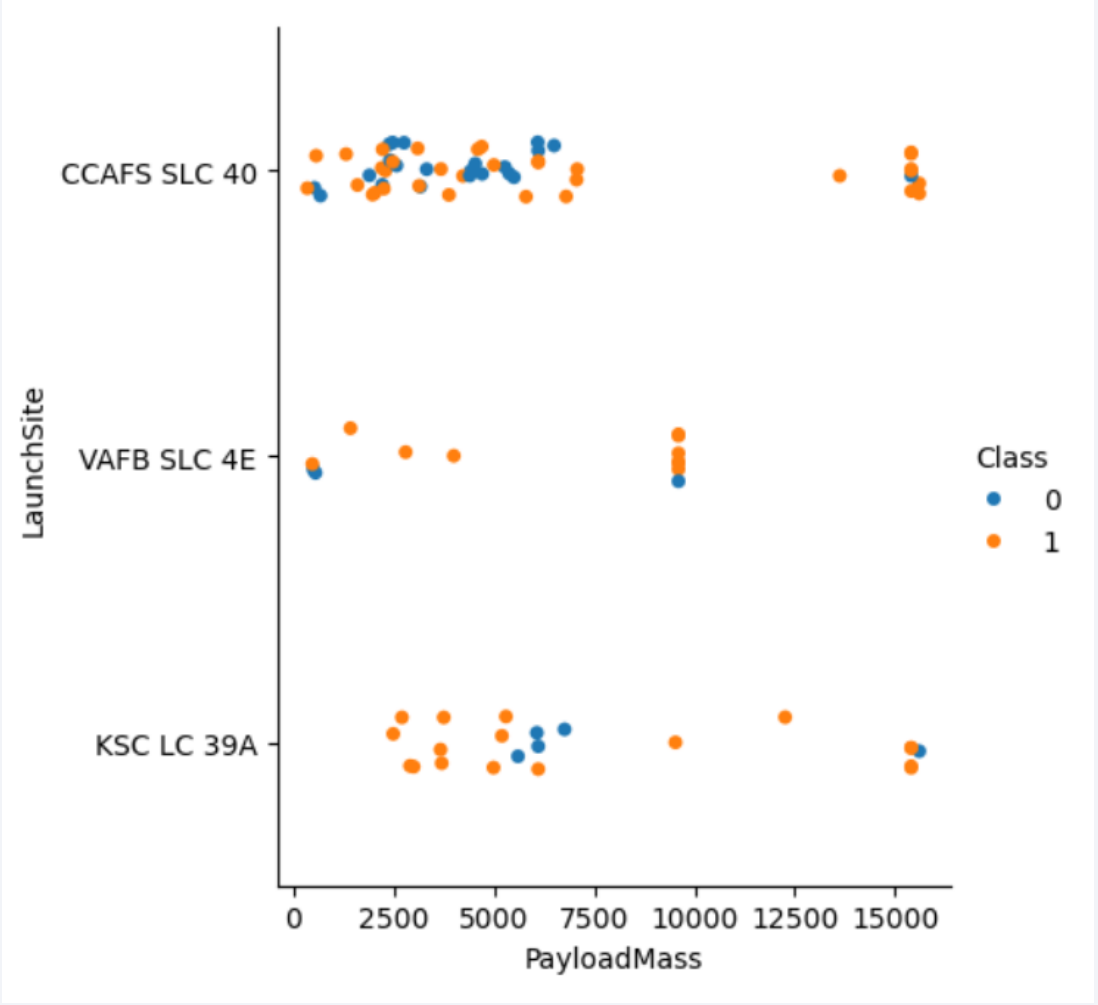
Flight Number vs. Launch Site

- The landing will succeed more likely as the number of flights increases



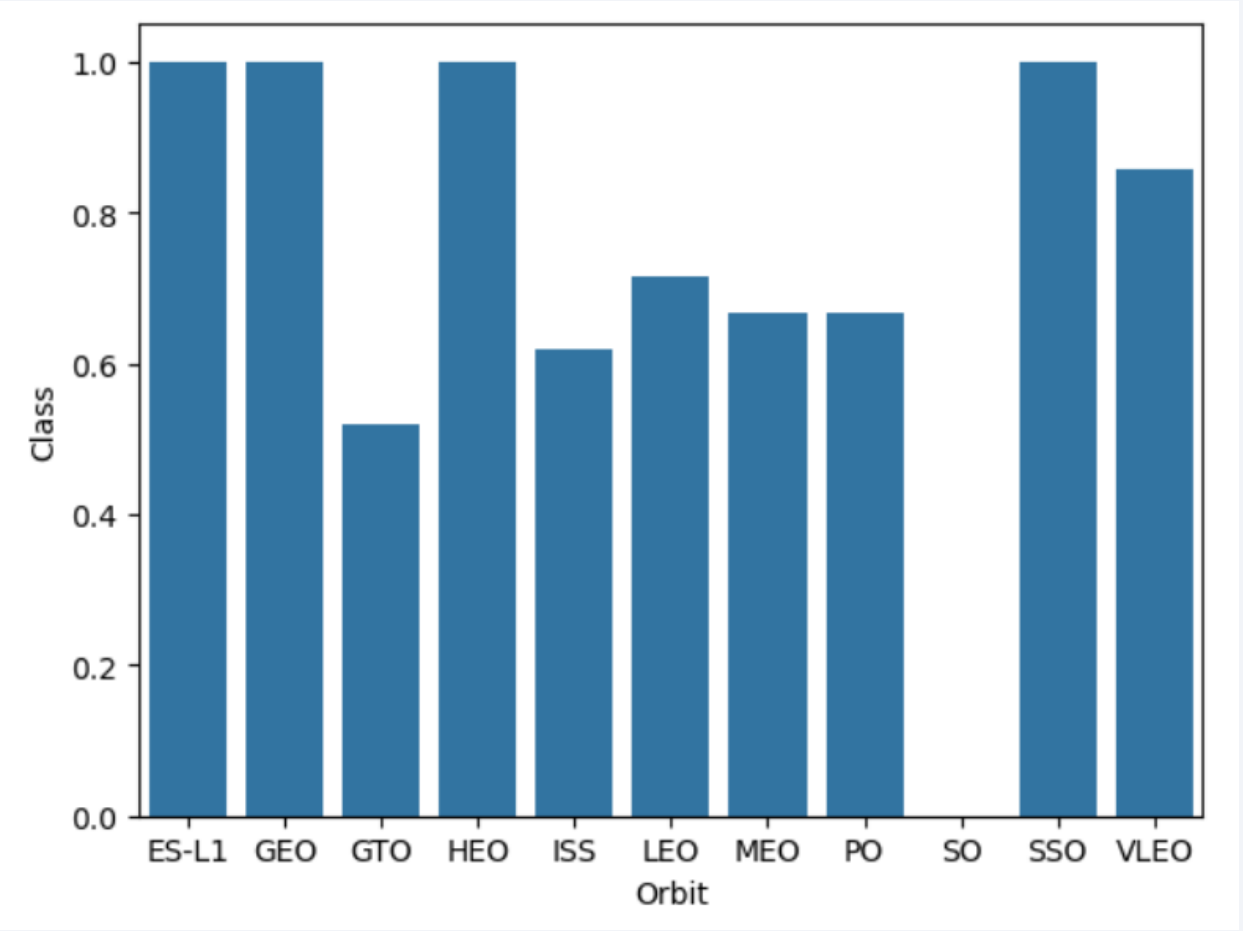
Payload vs. Launch Site

- Success rate is higher for CCAFS SLC 40 launch site with heavy payloads
- Over 10 000 kg payloads have not been used in VAFB SLC 4E launch site



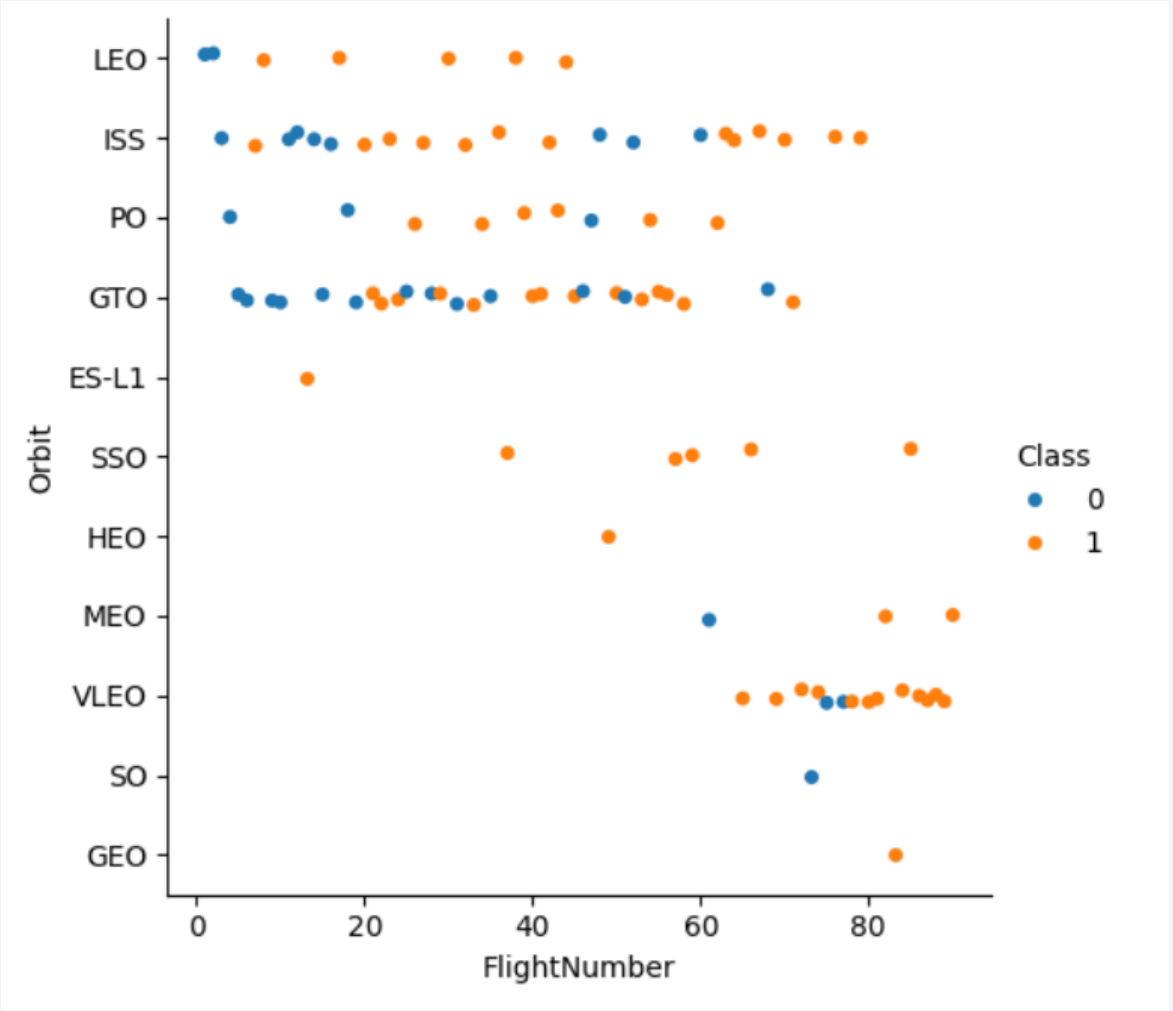
Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO and SSO have perfect success rates
- Orbit SO has zero success rate



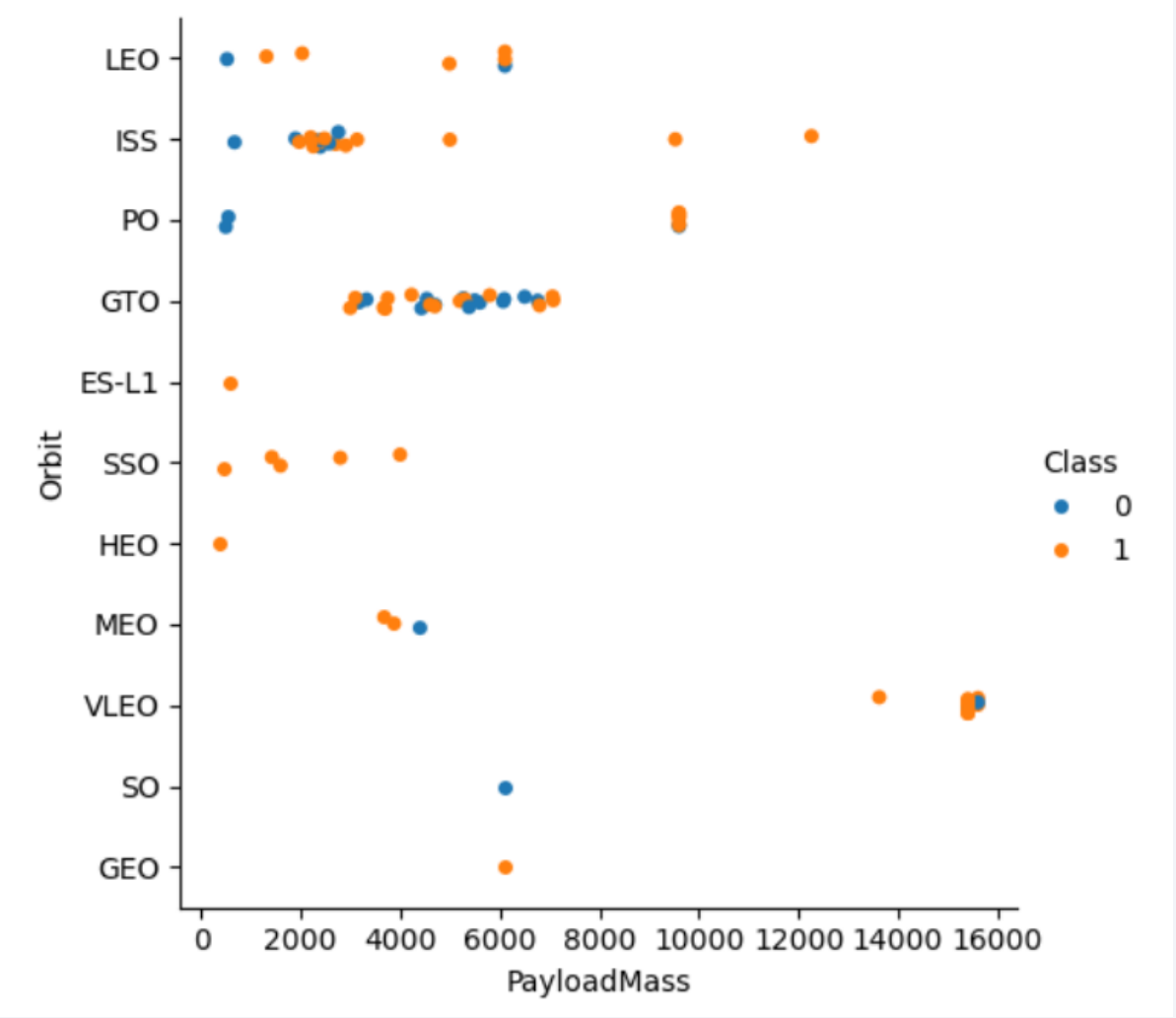
Flight Number vs. Orbit Type

- The success appears to be related to the number of flights in LEO orbit
- Other orbits do not have so clear relationship between flight number and success



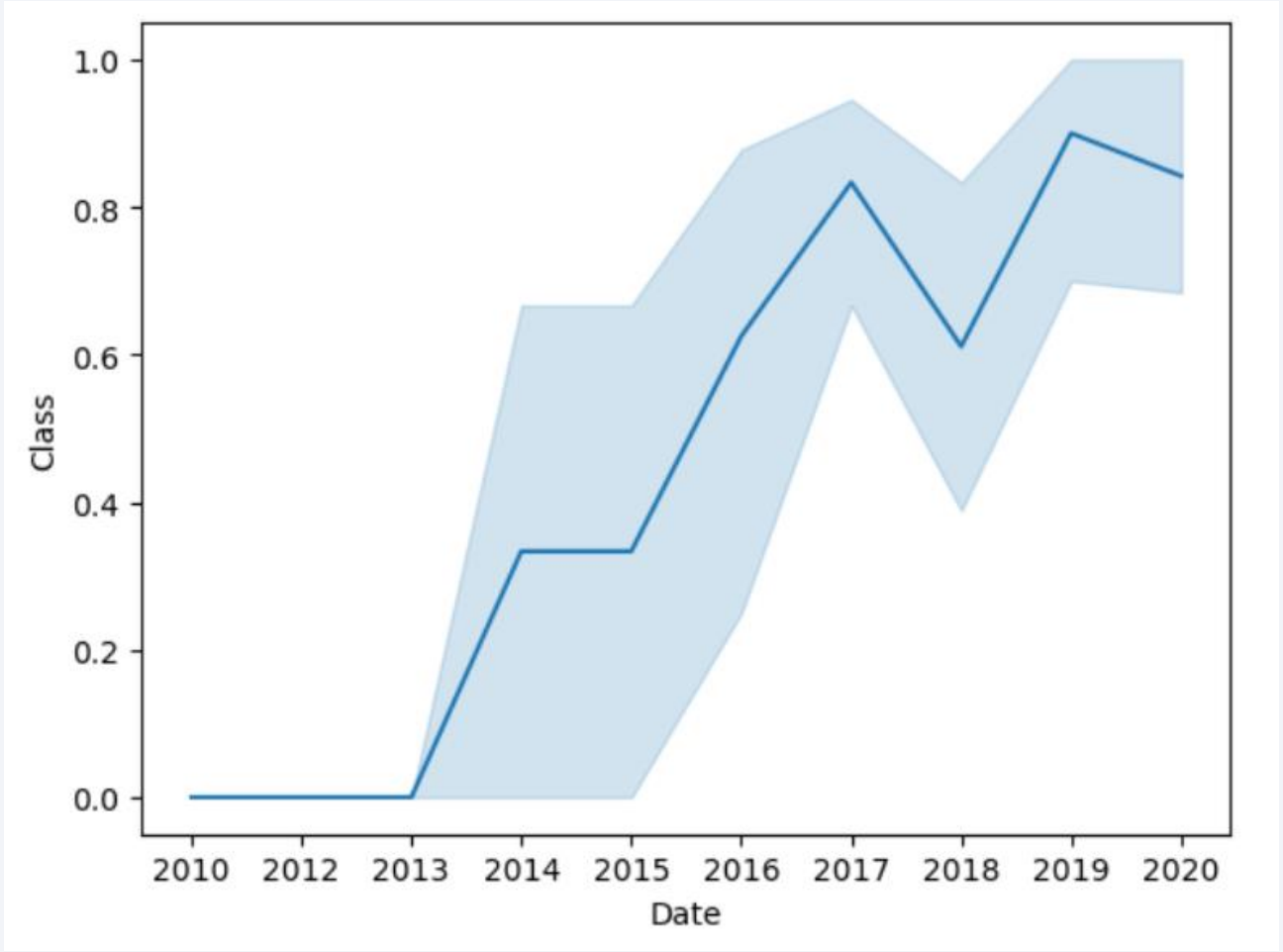
Payload vs. Orbit Type

- With heavy payloads the successful landing is more likely with orbits PO, LEO and ISS



Launch Success Yearly Trend

- The success rate increased from 2013 until 2019



All Launch Site Names

- SQL query of the names of the unique launch sites in the space mission

```
In [7]: %sql SELECT distinct(Launch_Site) FROM SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
```

Out[7]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- SQL query of 5 records where launch sites begin with the string 'CCA'

```
In [13]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE "CCA%" LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

Out[13]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (paracl
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (paracl
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No atte
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No atte
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No atte

Total Payload Mass

- SQL query of the total payload mass carried by boosters launched by NASA (CRS)

```
In [15]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = "NASA (CRS)";
 * sqlite:///my_data1.db
Done.
Out[15]: SUM(PAYLOAD_MASS__KG_)
                          45596
```

Average Payload Mass by F9 v1.1

- SQL query of the average payload mass carried by booster version F9 v1.1

```
In [16]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = "F9 v1.1";
 * sqlite:///my_data1.db
Done.
Out[16]: AVG(PAYLOAD_MASS__KG_)
         2928.4
```

First Successful Ground Landing Date

- SQL query of the date when the first successful landing outcome in ground pad was achieved

```
In [19]: %sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome LIKE "%Controlled%" or Landing_Outcome LIKE "%Success%";
 * sqlite:///my_data1.db
Done.
Out[19]: MIN(Date)
         2013-09-29
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- SQL query of the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
In [23]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = "Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND
```

```
 * sqlite:///my_data1.db
Done.
```

Out[23]:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- SQL query of the total number of successful and failure mission outcomes

```
In [26]: %sql SELECT Mission_Outcome, COUNT(*) FROM SPACEXTABLE GROUP BY Mission_Outcome;
 * sqlite:///my_data1.db
Done.
```

Out[26]:

Mission_Outcome	COUNT(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- SQL query of the names of the booster versions which have carried the maximum payload mass

```
In [28]: %sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FRO
 * sqlite:///my_data1.db
Done.
```

Out[28]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- SQL query of the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015.

In [38]:
```
%sql SELECT substr(Date, 4, 2), Landing_Outcome, Booster_Version, launch_site FROM SPACEXTABLE WHERE Landing_Outcome = "Fail
```

```
 * sqlite:///my_data1.db
Done.
```

Out[38]:

substr(Date, 4, 2)	Landing_Outcome	Booster_Version	Launch_Site
5-	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
5-	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- SQL query of the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

```
In [42]: %sql SELECT Landing_Outcome, COUNT(*) FROM SPACEXTABLE WHERE DATE BETWEEN "2010-06-04" and "2017-03-20" GROUP BY Landing_Out
 * sqlite:///my_data1.db
Done.
```

Out[42]:

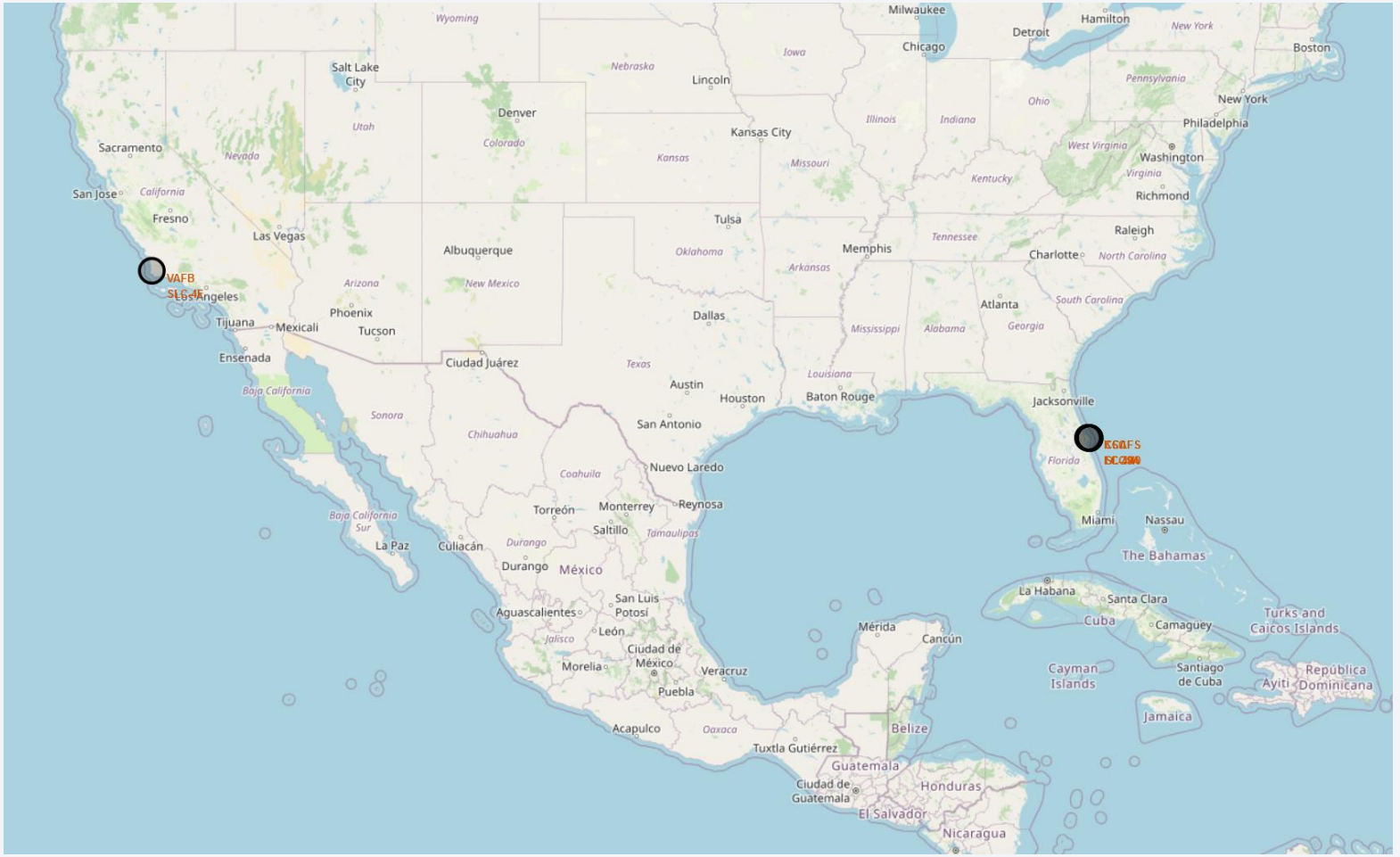
Landing_Outcome	COUNT(*)
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

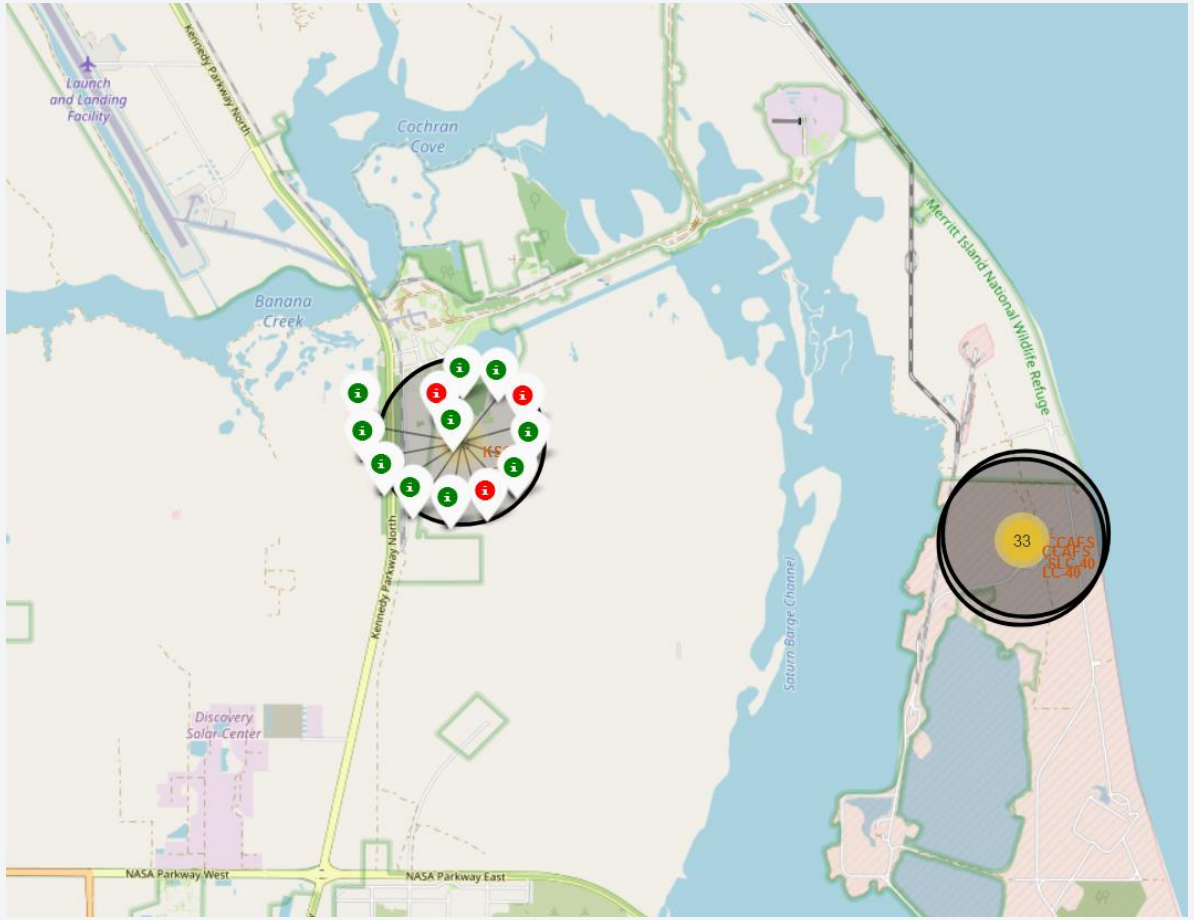
Launch site locations

- All launch sites are located near coastline



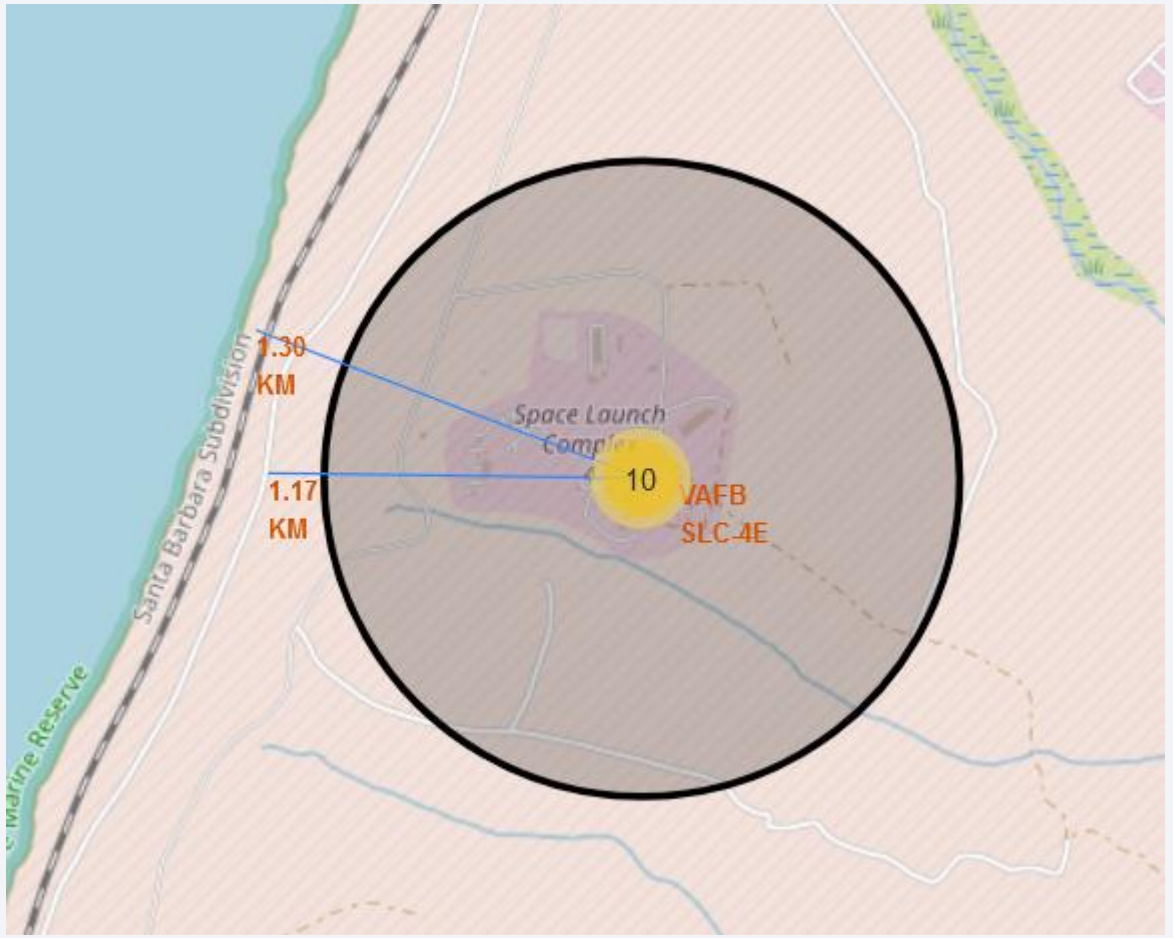
Launch outcomes in KSC LC-39A launch site

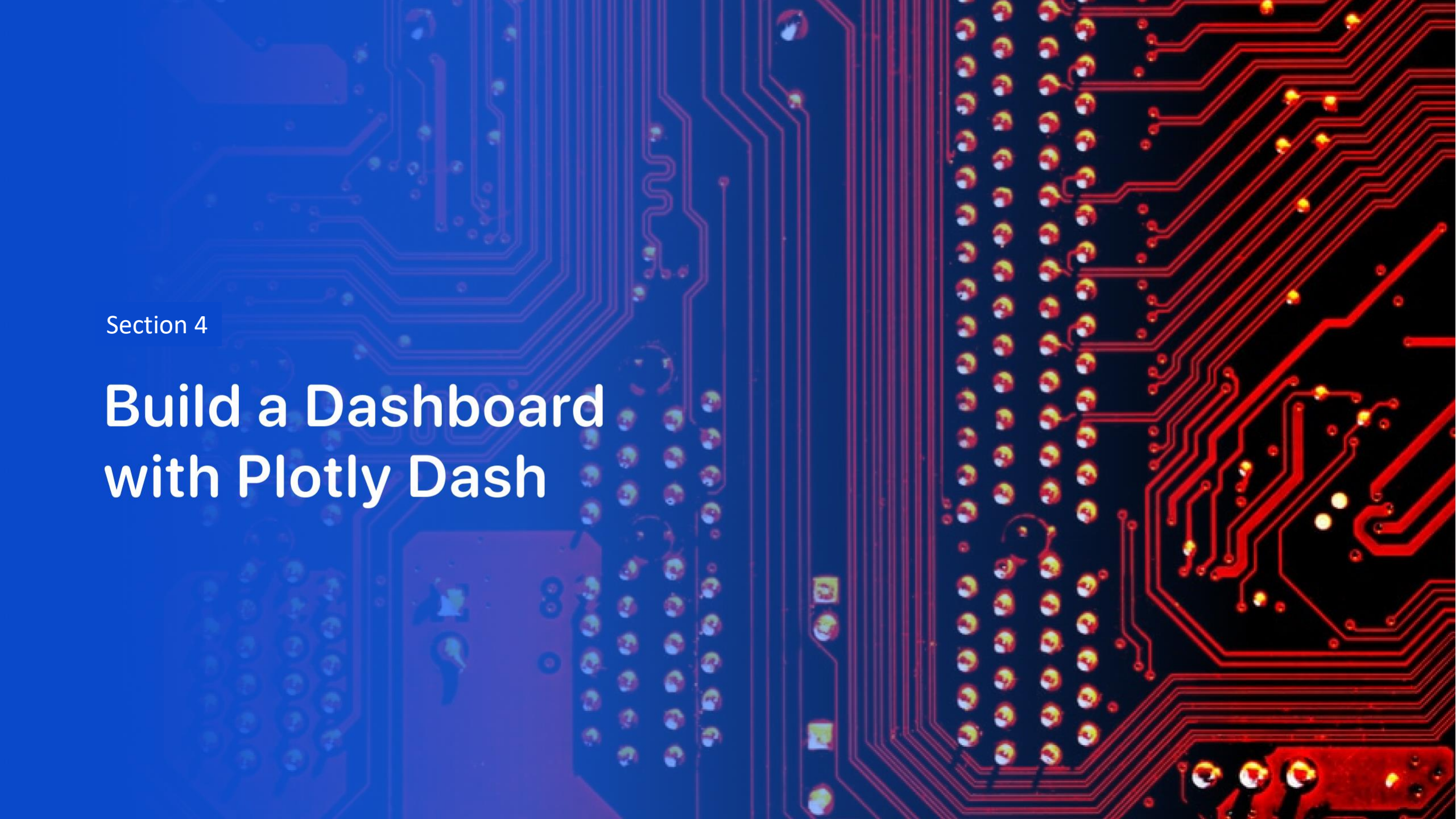
- Green marker = successful launch
- Red marker = failed launch
- KSC LC-39A has the highest success rate



Distances between VAFB SLC-4E and landmarks

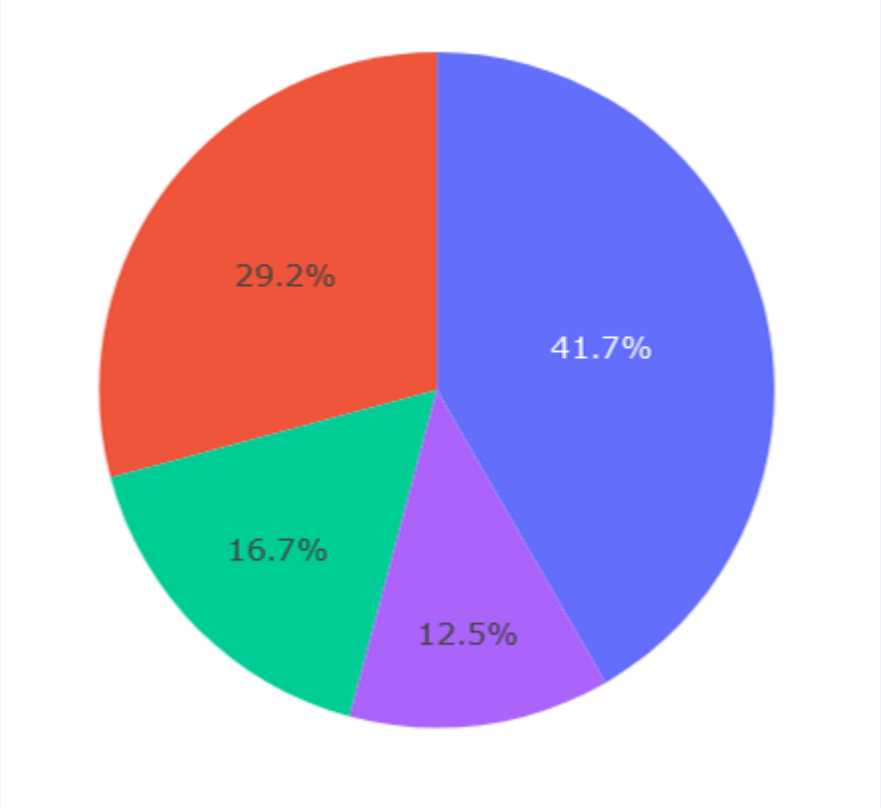
- Distances from launch site VAFB SLC-4E to important landmarks in its proximities
  - To coastline 1.3 km
  - To coast road 1.17 km




Section 4
Build a Dashboard with Plotly Dash

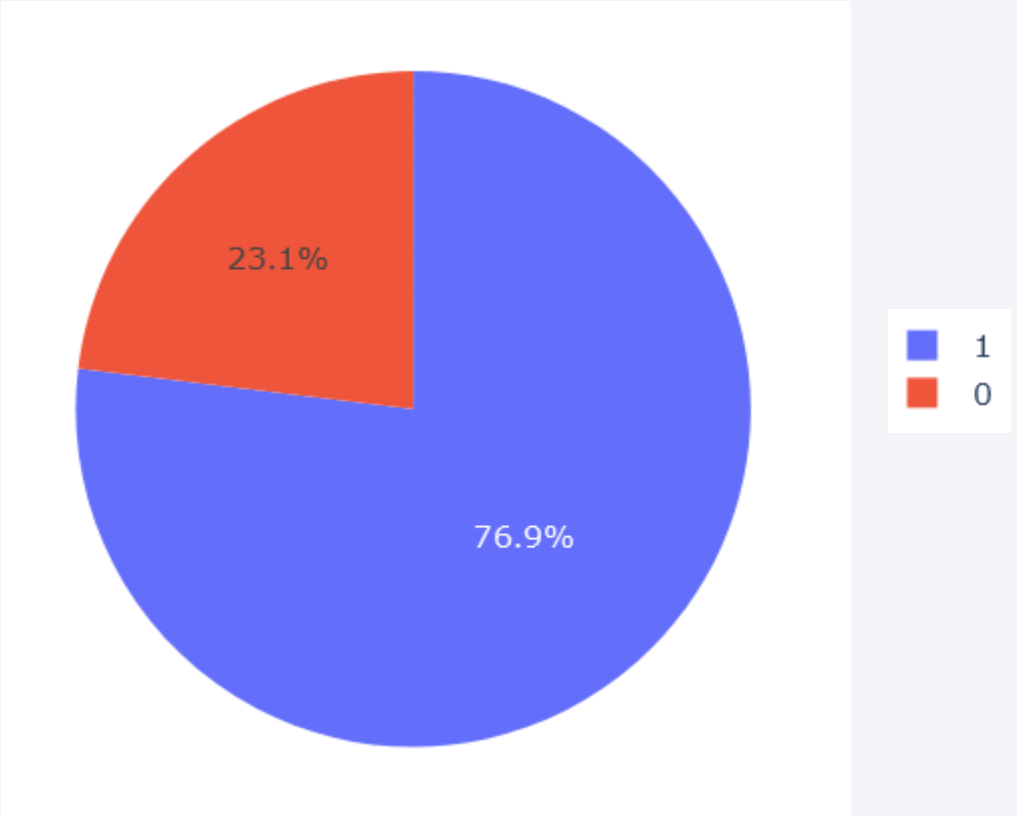
Launch success count for all sites

- KSC LC-39A has the highest launch success count
- CCAFS SLAC-40 has the lowest launch success count



Launch success ratio in KSC LC-39A launch site

- KSC LC-39A launch site has the highest success ratio with 10 successful and 3 failed launches



Payload vs. Launch Outcome with different payloads

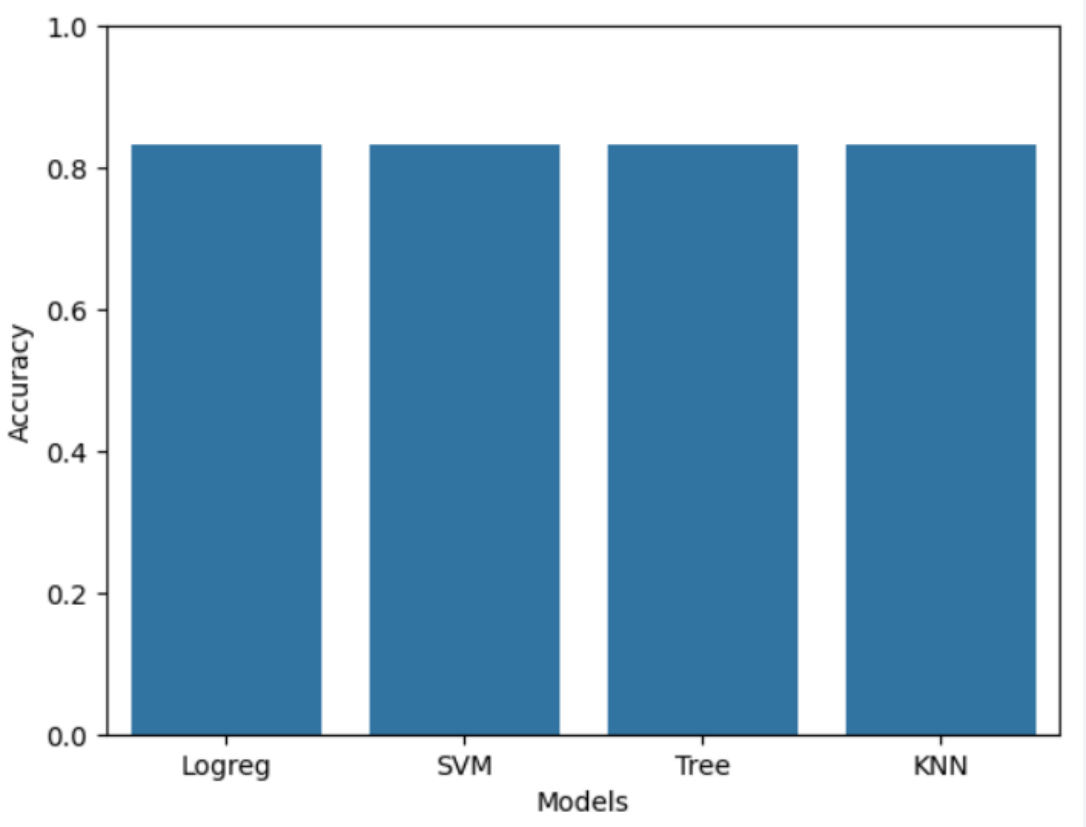
- Payloads at the range of 2000 – 5500 kg have the highest success rate

Section 5

Predictive Analysis (Classification)

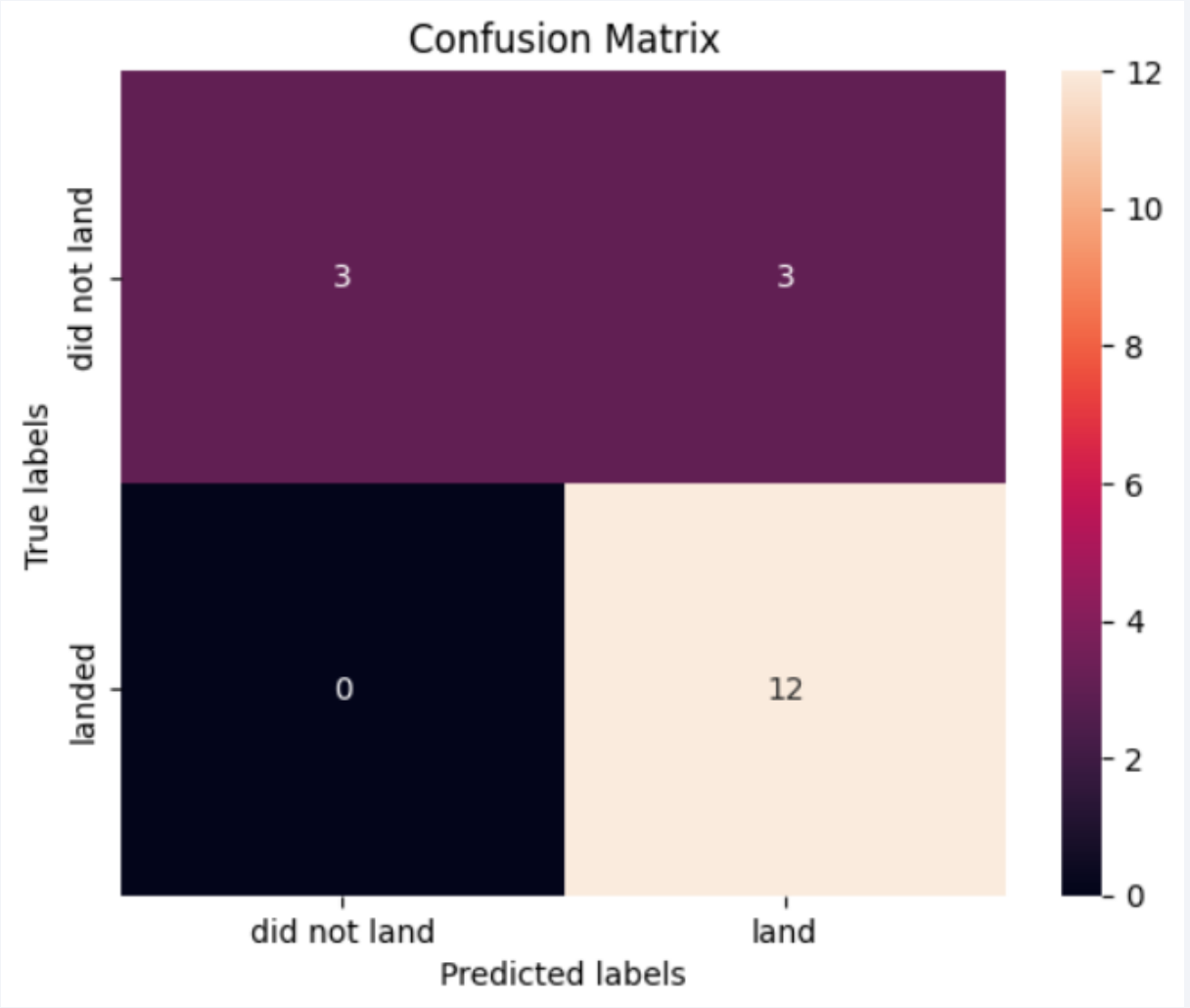
Classification Accuracy

- All models give the accuracy of 0.83 for the test set
- Test set contains only 18 samples



Confusion Matrix

- All models provided the same confusion matrix
- Models can predict perfectly all successful landings
- The major problem is that models provide only 50 % accuracy for unsuccessful landings



Conclusions

- Orbits ES-L1, GEO, HEO and SSO have perfect success rates
- All launch sites are located near coastline
- Launch site KSC LC-39A has the highest success rate
- Logistic regression, SVM, Tree and KNN classification algorithms all provide same accuracy (0.83) and confusion matrix
- Models can predict correctly all successful landings but only 50 % of unsuccessful landings

Thank you!